The data supplied in the file `air_pollution.txt` are on air pollution in 41 U.S. cities. The type of air pollution under study is the annual mean concentration of sulfur dioxide. It is desired to develop a regression model to predict air pollution using 6 explanatory variables \( x_1 - x_6 \) measured as described on the back of this page. A SAS procedure was used to fit a multiple regression model:

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon
\]

using data for all 41 cities.

1. Use a scatterplot matrix and correlation matrices to carry out a preliminary assessment of the pairwise relationships between \( y \) and \( x_1, x_2, x_3, x_4, x_5, \) and \( x_6. \) On this basis alone, select a few explanatory variables that may good predictors in a multiple regression model. Using the plot and the correlation matrix, find the four variables that are most strongly correlated among the explanatory variables. Based on above analysis alone suggest the explanatory variables that are most strongly involved in multicollinearity when fitting the full model.

2. Use a SAS procedure to fit the above multiple regression model to all 41 cities. Discuss the fit of this model using the Anova table, \( R^2 \), and the estimates table. Use other diagnostic tools including output statistics and plots of residuals to examine the adequacy of the model (use the diagnostic panel of plots). Comment on these cases, identifying any possible problems with model assumptions indicated by them and cases that may be outliers and/or influential.

3. Identify any least squares estimates of the regression coefficients (\( \hat{\beta} \)'s) from the fit of the model in Prob 2.), that have a sign (positive or negative) that is different from what you would expect for the parameter – an indication of multicollinearity? Use the standard errors of the parameter estimates to show that these are poorly estimated. Do the variance inflation factors (VIF’s) identify these parameters?

4. Remove the explanatory variables \( x_3 \) and \( x_6 \) from the five-variable model and use a multiple regression model to relate \( y \) to \( x_1, x_2, x_4 \) and \( x_5 \) only. What can you observe about the multicollinearity in the new model? Is there improvement in the accuracy of estimation of parameters of this model (for e.g., decreases standard errors, more t-statisticas are significant etc.)? Justify your answers.

5. Perform a residual analysis using diagnostic statistics including relevant graphics for the above 4-variable model. Specifically, determine whether a single city does not appear to fit the model very well and provide evidence from your analysis for making this conclusion.

6. Remove the case you determined above to be a possible outlier from the data and use all 6 variables for the analyses described in the following three parts:

   (i) Use a SAS procedure to do all possible regressions containing no less than 2 and no more than 4 explanatory variables. Print statistics for only the 4 best models in each case. Construct a plot of the \( C_p \) statistic for all models with ‘reasonable’ \( C_p \) values. Select a single model, each with 2, 3, and 4 explanatory variables, respectively, for the purpose of predicting annual mean concentration of sulfur dioxide in a city, indicating your reasons for selection of each model. There may be several possible choices, i.e., there may be many ‘good’ models but give arguments for each of your choices. Primarily, use \( s^2, R^2, \) and \( C_p \) in your arguments. Select one of these models as your final model and provide arguments supporting your choice.
(ii) Use the backward elimination subset selection procedure with significance level of 0.05 for deleting variables to select a possible model. State the model selected and report estimates of parameters and the analysis of variance table for this model.

(iii) Use the stepwise subset selection procedure, with significance levels of 0.10 for entry and 0.05 for deletion of variables, respectively, to select a possible model. State the model selected and report estimates of parameters and the analysis of variance table for this model.

Data: Note that the first column of the data set is a city number. Data values for the variables appear in the order they are described below. A link to a text file is available on the assignment page.

city = city number (input as character data)
y = the annual mean concentration of sulfur dioxide (mcg/meter$^3$)
x1 = average annual temperature °F
x2 = number of manufacturing enterprises employing 20 or more workers
x3 = population size (1970 census) in thousands
x4 = average annual wind speed (mph)
x5 = average annual precipitation (inches)
x6 = average number of days with precipitation per year

Note: It may take several computer runs to complete the required analyses. You may cut-and-paste parts of computer printout into your answer sheets.

Due Tuesday December 1, 2015